

Set	Items	Description
S1	35	ENTROPY (S) CLUSTER?
S2	31	RD S1 (unique items)
S3	1	S2 AND SHANNON

3/3,K/1 (Item 1 from file: 148)  
 DIALOG(R) File 148:Gale Group Trade & Industry DB  
 (c)2004 The Gale Group. All rts. reserv.

16674446 SUPPLIER NUMBER: 111300301 (USE FORMAT 7 OR 9 FOR FULL TEXT  
 )

**Finding the number of clusters in a dataset: an information-theoretic approach.**

Sugar, Catherine A.; James, Gareth M.

Journal of the American Statistical Association, 98, 463, 750 (14)

Sept, 2003

ISSN: 0162-1459 LANGUAGE: English RECORD TYPE: Fulltext; Abstract

WORD COUNT: 11584 LINE COUNT: 00960

... distortion between the source,  $X$ , and its representation, (  
 $.X)$ , and  $I(X; ($   
 $.X))$  is the Shannon mutual information between  $X$  and (  
 $.X)$ . The mutual information is defined as  
 $I(X; ($   
 $.X...$

...the information-theoretic statistics literature and its relationship to  
 the pioneering work of C. E. Shannon was given by Soofi (1994).

#### 2.2 Asymptotic Rate Distortion Theory Results

Here we give...

...the distribution of  $X$ .

The first result suggests that any choice of the number of clusters  
 based on the distortion curve or monotone transformations thereof will be  
 admissible in the sense...

...to establish. Sugar (1999) gave a proof of convexity under certain  
 hierarchical restrictions on the clustering methodology. Results (II) and  
 (III) follow from the maximum entropy property of the Gaussian. Versions  
 of (II) exist for more-complex covariance structures. However, it...

...of clusters.

Most of the fundamental work in this area is due to C. E. Shannon ,  
 who pioneered the field of mathematical communication ( Shannon 1948) and  
 introduced the notion of a rate distortion function ( Shannon 1959). Cover  
 and Thomas (1991) present a more complete development, including extensive  
 references and proofs...Y. T. (1985), "A Criterion for Determining the  
 Number of Clusters in a Data Set," Biometrics , 44, 23-34.

Kullback, S., and Leibler, R . A. (1951), "On Information and  
 Sufficiency," The Annals of Mathematical Statistics, 22, 79-86.

Lindley, D. V. (1956), "On a Measure of the Information Provided by  
 an Experiment ," The Annals of Mathematical Statistics, 27, 986-1005.

McEliece, R . J. (1977), The Theory of Information and Coding: A  
 Mathematical Framework for Communication, Reading, MA: Addison-Wesley.

Milligan

?